Groupe outlier factor et détection de nouveautés

Amine Chaibi

Vichy, 2 juin 2014

Contexte général : Cifre LIPN et Anticipeo

1. B.D difficiles : données manquantes, "outliers", déséquilibre...
2. Définir les produits et les clients homogènes ;

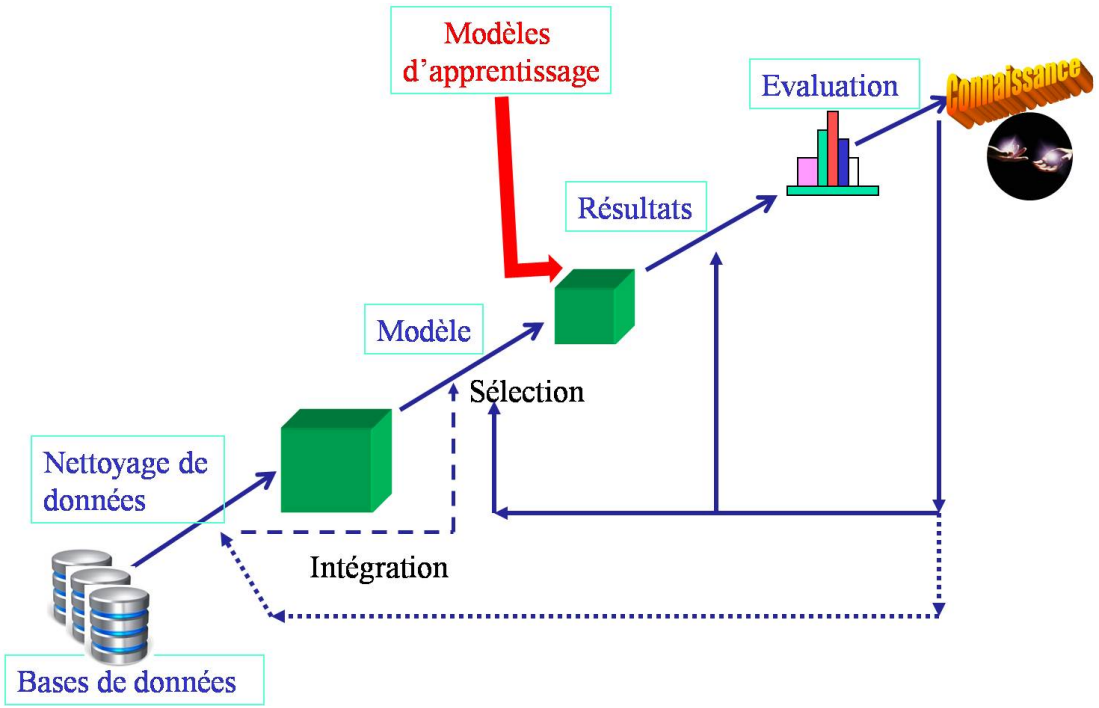
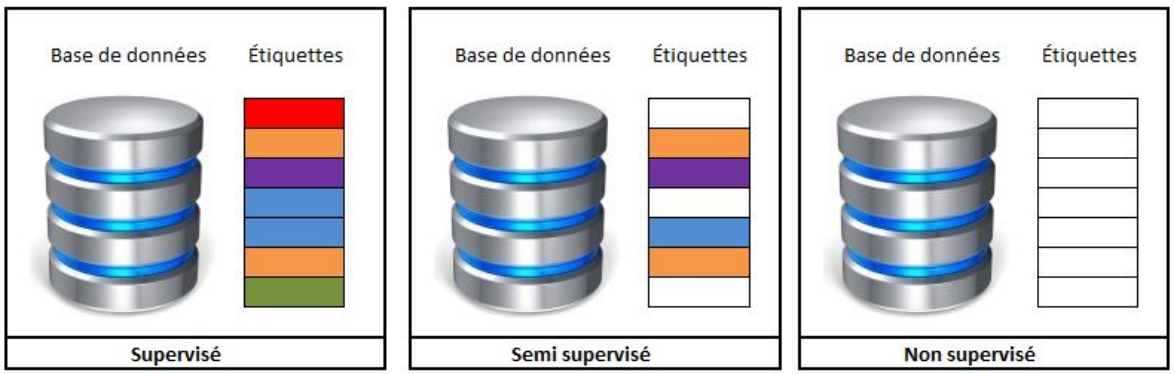
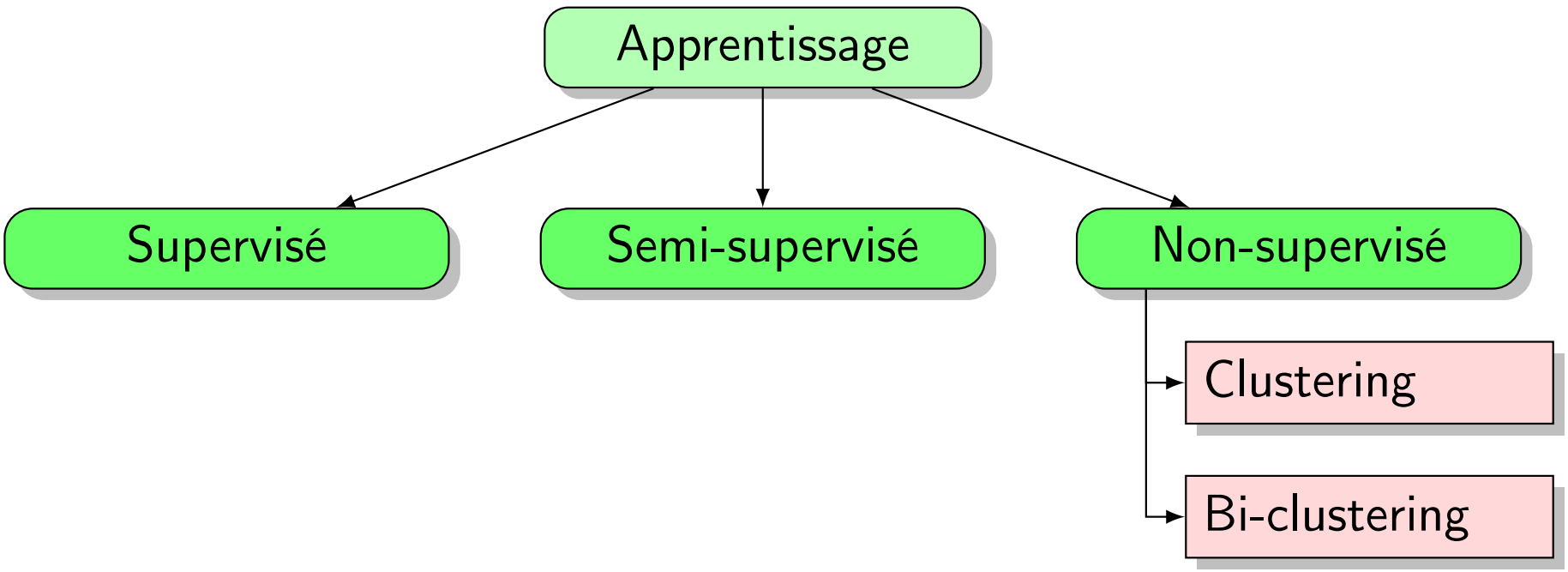


Figure : Processus data mining

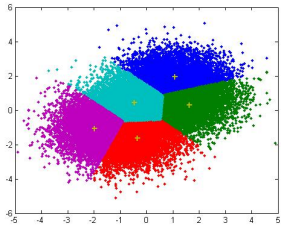
Paradigme d'apprentissage : déterministe/probabiliste

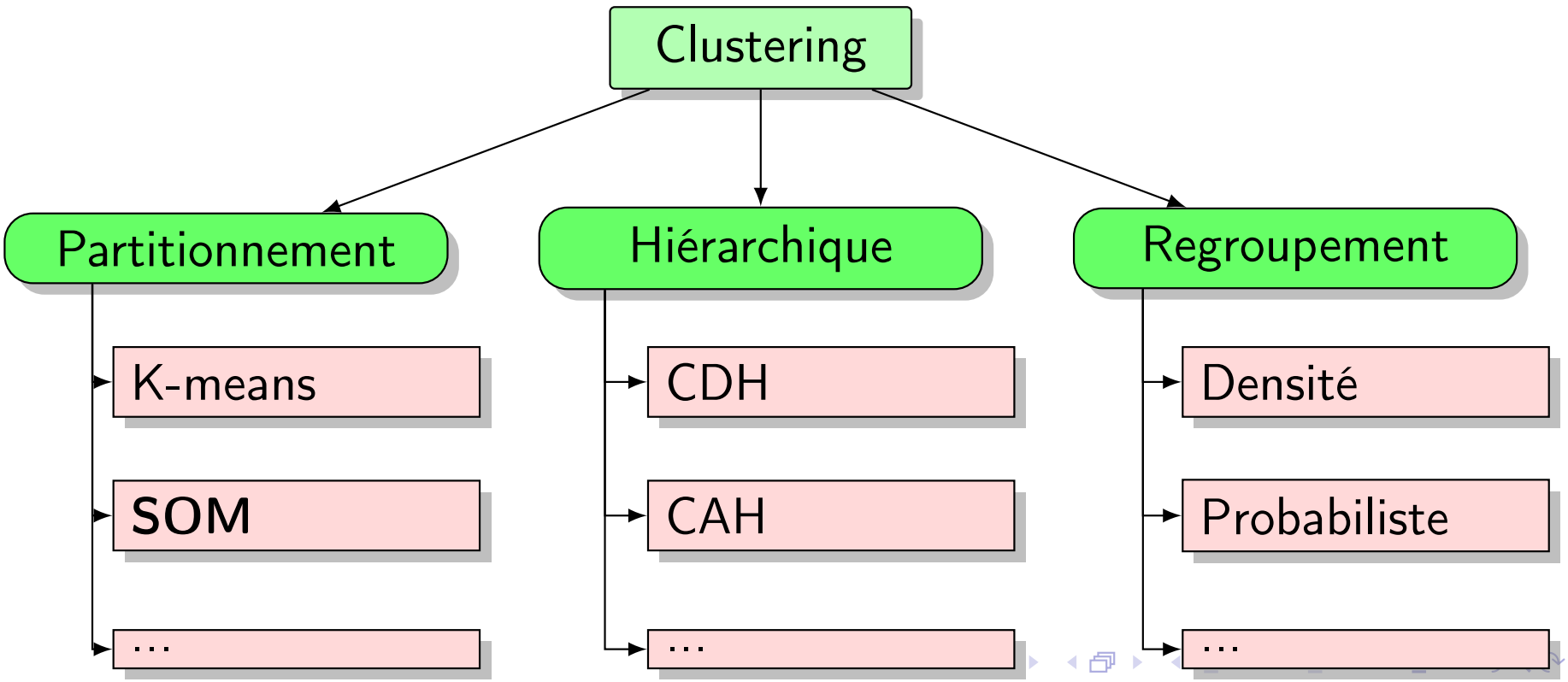




Clustering [Cormack 71]

- Regrouper les observations d'un ensemble de données en classes homogènes.





Définition

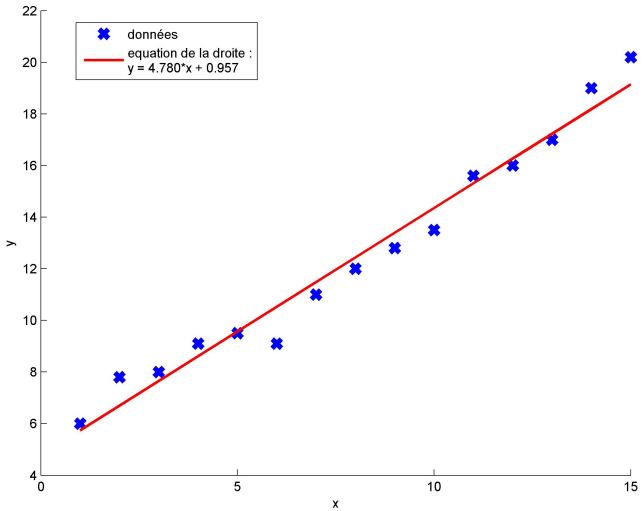
Peirce 1852 : *"dans presque toutes les séries de données, il y a des observations qui diffèrent tellement des autres, qu'elles servent uniquement à rendre l'expérimentateur perplexe et à l'induire en erreur !"*

Définition

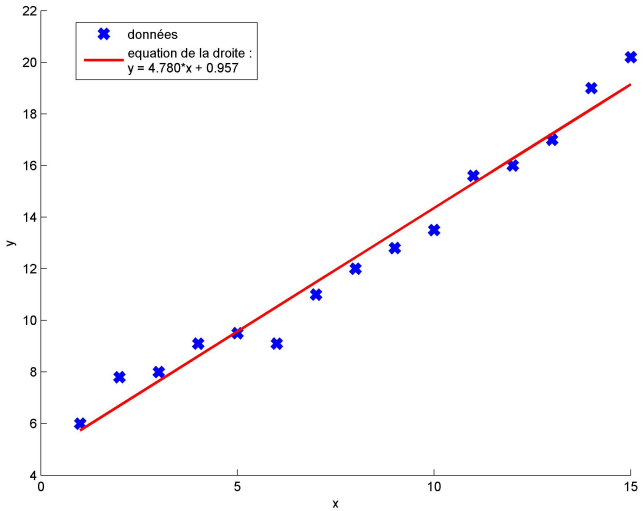
Peirce 1852 : *"dans presque toutes les séries de données, il y a des observations qui diffèrent tellement des autres, qu'elles servent uniquement à rendre l'expérimentateur perplexe et à l'induire en erreur !"*

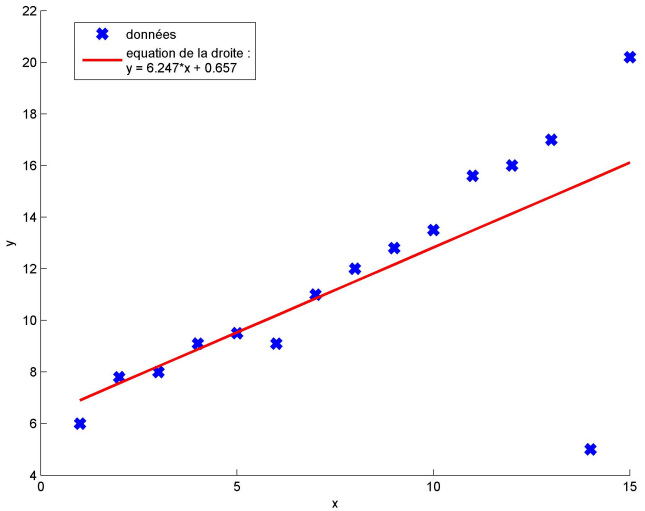
- Un "**outlier**" est une observation qui n'est pas conforme ou normale par rapport au comportement global de l'ensemble des données.
- Un "**groupe-outlier**" est un ensemble de données formant un groupe dense et significativement isolé.
- Une "**nouveauté**" est une donnée qui n'était pas connue dans la base d'apprentissage et qui apparaît dans la base de test.

Les outliers et la régression

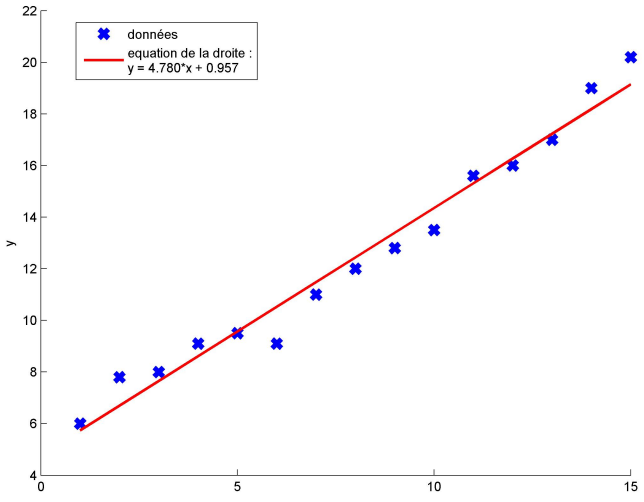


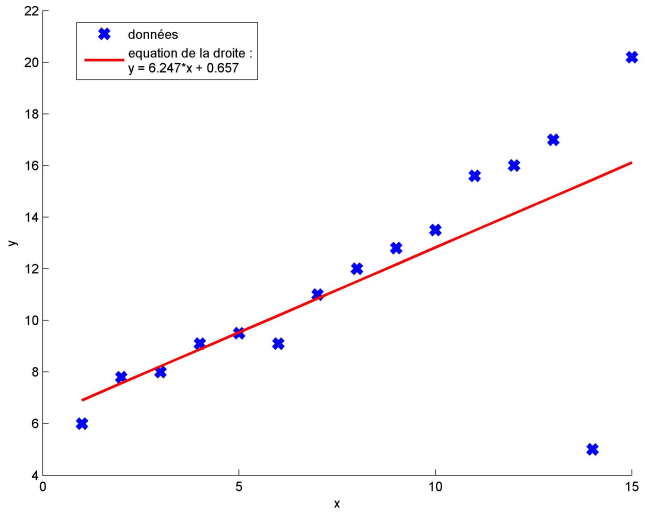
Les outliers et la régression

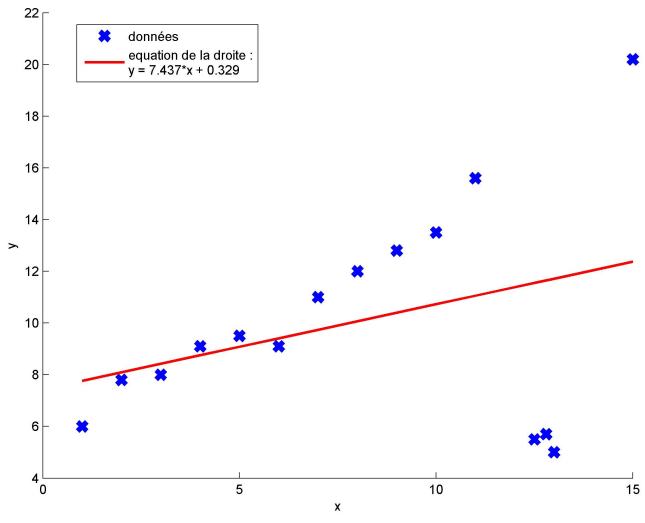




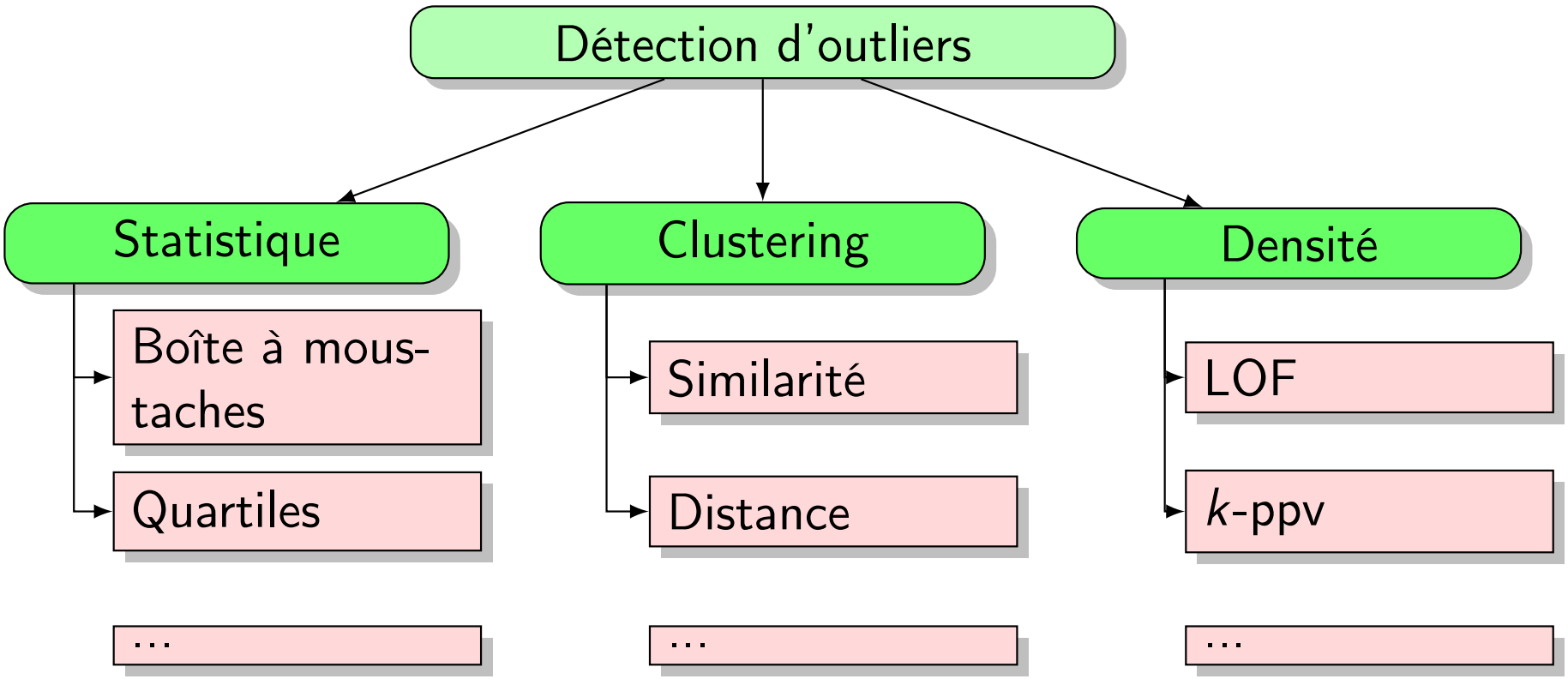
Les outliers et la régression



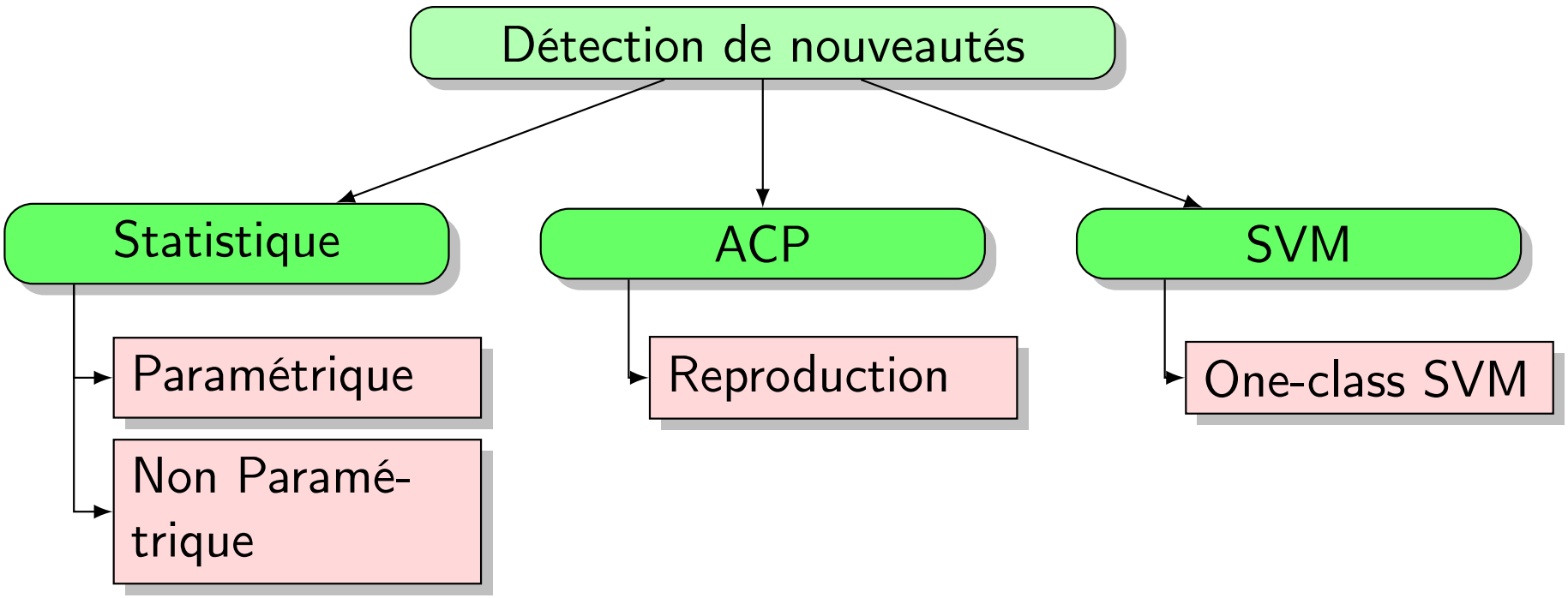




Détection d'outliers



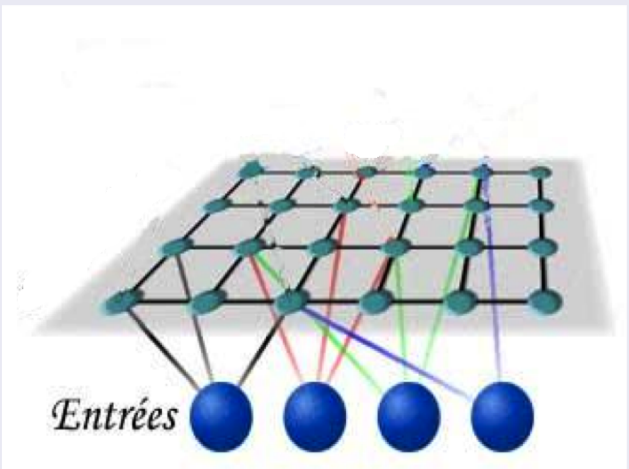
Détection de nouveautés



Evaluation : rappel (tvp), ROC, ...

Cartes SOM [Kohonen 1995]

Grille possédant un ordre topologique de K cellules.



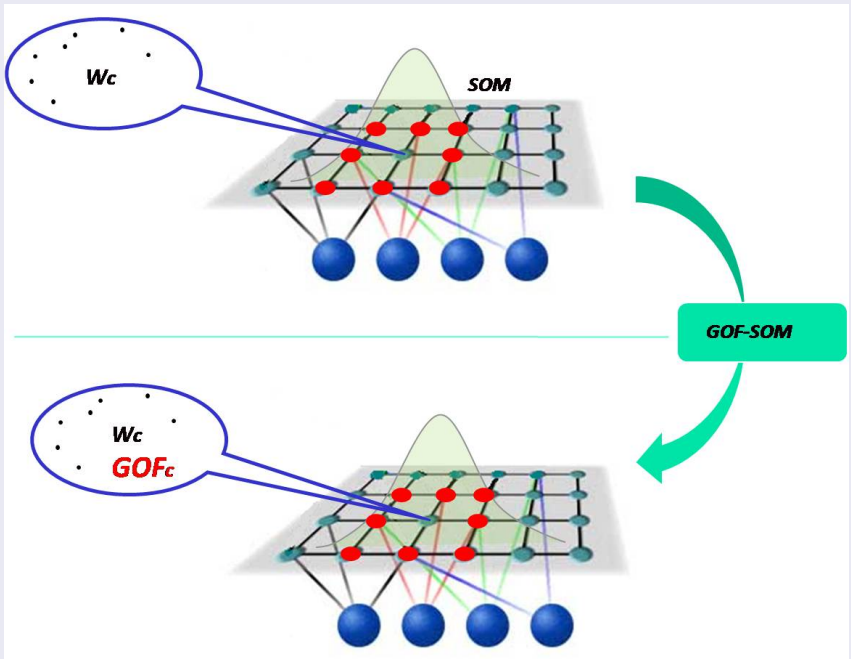
Modèle SOM

1. Batch : affectation de la **matrice** puis minimisation ;
2. Stochastique : affectation d'une **donnée** puis minimisation.

$$\mathcal{J}_{SOM}(\mathcal{W}, \phi) = \sum_{i=1}^{N} \sum_{c=1}^{K} K^T(\delta(\phi(x_i), c)) d(w_c - x_i)$$

Les "outliers" et les cartes topologiques

- Estimation d'un score "d'outlier-ness" pour chaque référent de la carte.
- Détection automatique des "groupe-outliers".



GOF intégré aux cartes topologiques [ESANN 2012]

$$Min\mathcal{R}(\mathcal{W}, GOF, \phi) = \mathcal{R}(\mathcal{W}, \phi) + \mathcal{R}(GOF, \phi)$$

où

$$\mathcal{R}(\mathcal{W}, \phi) = \sum_{i=1}^{N}\sum_{c=1}^{K} K^T(\delta(\phi(x_i), c))\|w_c - x_i\|^2)$$

et

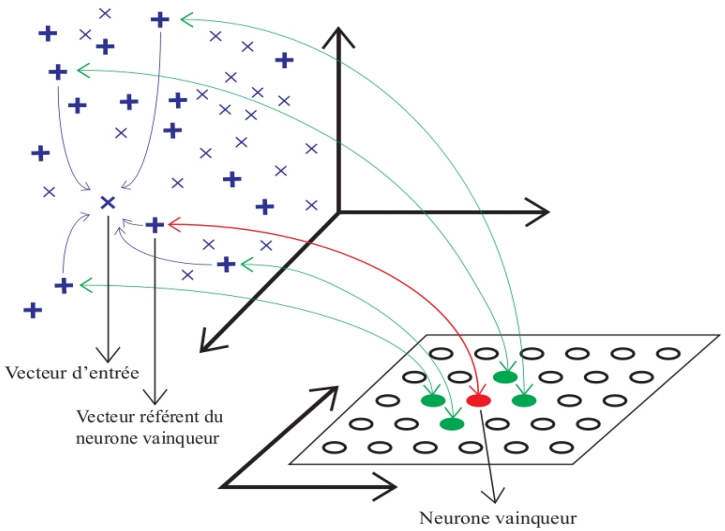
$$\mathcal{R}(GOF, \phi) = \sum_{i=1}^{N}\sum_{c=1}^{K} K^T(\delta(\phi(\mathbf{x}_i), c))(GOF_c - OF_c(x_i))^2$$

- $OF_c(\mathbf{x}_i) = \dfrac{\dfrac{\sum_{\mathbf{x}_j \in P_c} \frac{1}{f_c(\mathbf{x}_j)}}{|P_c|}}{\frac{1}{f_c(\mathbf{x}_i)}}$
- $f_c(x_i) = \exp^{-\frac{\|w_c - x_i\|^2}{2\sigma^2}}$ estime la densité (empirique).

Phase de compétition

Affecter une donnée x_i en utilisant la fonction

$$\phi(x_i) = \arg \min_{1 \leq j \leq K} \| x_i - w_j \|^2$$



Phase d'adaptation : version stochastique (descente de gradient)

- Mettre à jour les référents w_c de chaque cellule c :
  $w_c(t) = w_c(t-1) - \varepsilon(t) K^T(\delta(\phi(x_i), c))(w_c(t-1) - x_i)$
- Mettre à jour les valeurs de GOF_c associées à chaque cellule c :
  $GOF_c(t) = GOF_c(t-1) -$
  $$\varepsilon(t) K^T(\delta(\phi(\mathbf{x}_i), c)) \left(GOF_c(t-1) - \frac{\sum_{\mathbf{x}_j \in P_c} \frac{1}{f_c(\mathbf{x}_j)}}{\frac{|P_c|}{\frac{1}{f_c(\mathbf{x}_i)}}} \right)$$
  où $\varepsilon(t)$ est le pas d'apprentissage
- Répéter les deux phases jusqu'à $t = t_{max}$.

Applications

- Visualisation ;
- Détection de nouveautés.

Bases de données utilisées : OCC et UCI

Base de données	# Obs	# Var	# Normales	# Outliers
Iris Setosa	150	4	50	100
Iris Virginica	150	4	50	100
Sonar Mines	108	60	11	97
Biomed Healthy	194	5	127	67
Hepatitis Normal	155	19	123	32
Diabetes Present	768	8	500	268
Ecoli Periplasm	336	7	52	284
Spectf 1	349	44	254	95
Balance-Scale	625	4	288	337
Glass Building	214	9	70	144
Waveform 2	900	21	300	600

Nom de la base	# Obs	Taille	Nom de la base	# Obs	Taille
anneauxM	1072	14×12	demicercleM	638	13×10
HeptaM	212	9×8	LsunM	400	11×9
TargetM	951	13×12	GolfBallM	4343	19×17
base simulée 1	160	5×13	base simulée 2	234	3×26
base simulée 3	569	8×15	base simulée 4	402	8×13

Visualisation des groupes-outliers en utilisant GOF-SOM

[ICONIP 2012]

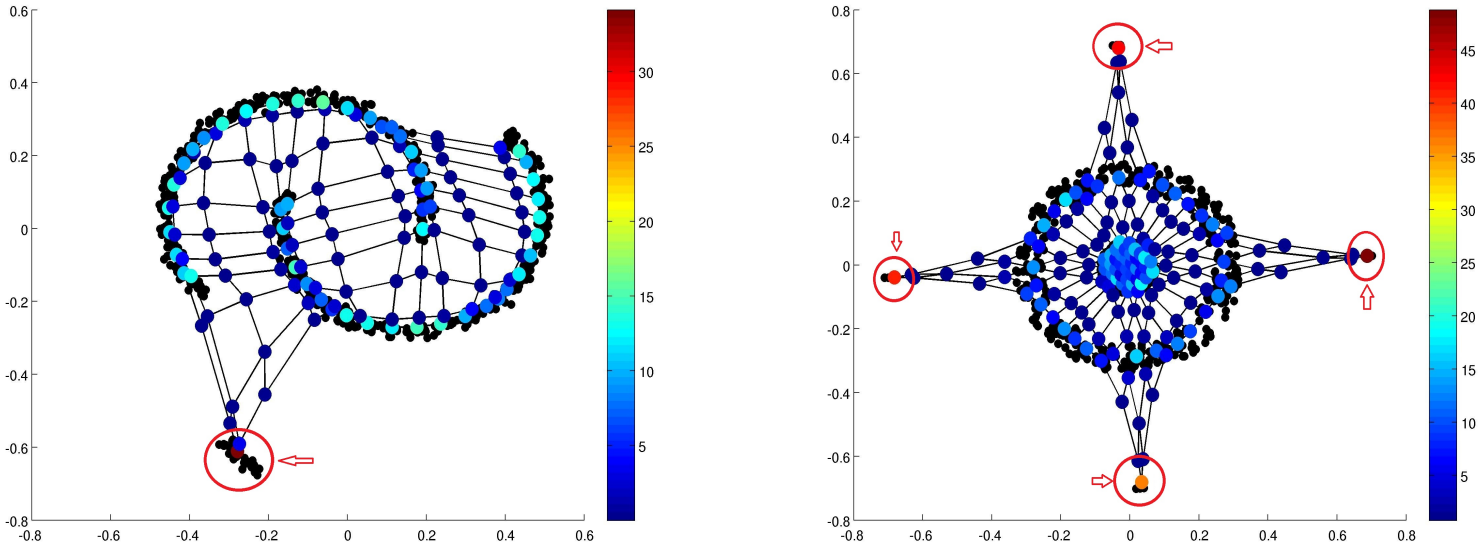
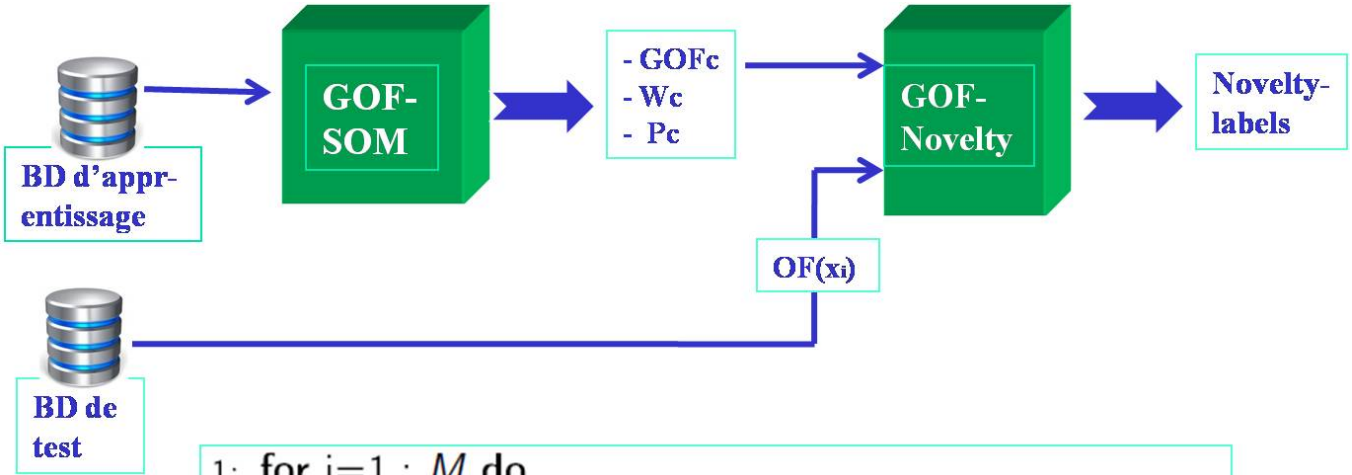


Figure : GOF-SOM appliqué sur les données des bases demicercleModif et TargetModif

Limite : taille de la carte.

GOF : détection de nouveautés [ICONIP 2012, IJCIA 2013]



1: for i=1 : M do

2: $OF_{\phi(x_i)}(x_i) = \frac{\frac{\sum_{x_j \in P_{\phi(x_i)}} \frac{1}{fc(x_j)}}{|P_{\phi(x_i)}|}}{\frac{1}{fc(x_i)}}$

3: $Dif = |OF_{\phi(x_i)}(x_i) - GOF_{\phi(x_i)}|$

4: if $(Dif < threshold)$ then

5: $Novelty\_label(x_i) = 0$;

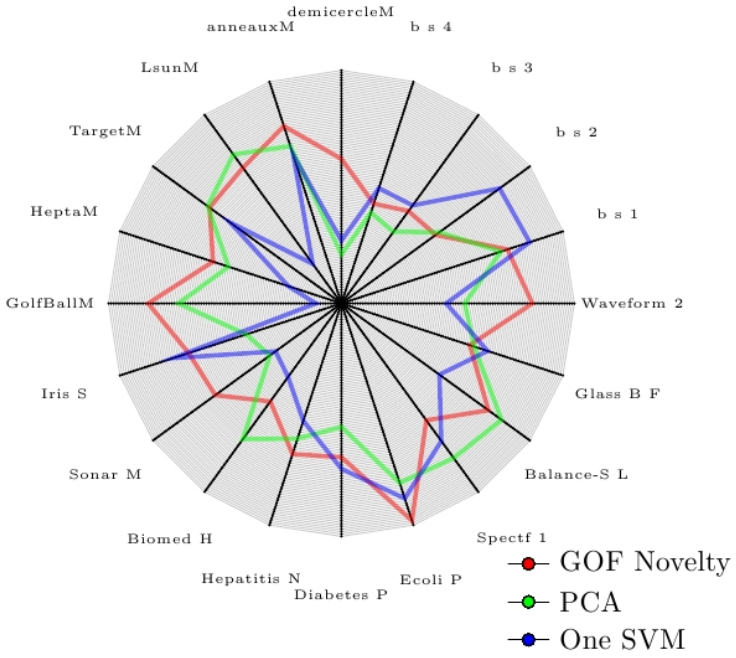
6: else

7: $Novelty\_label(x_i) = 1$;

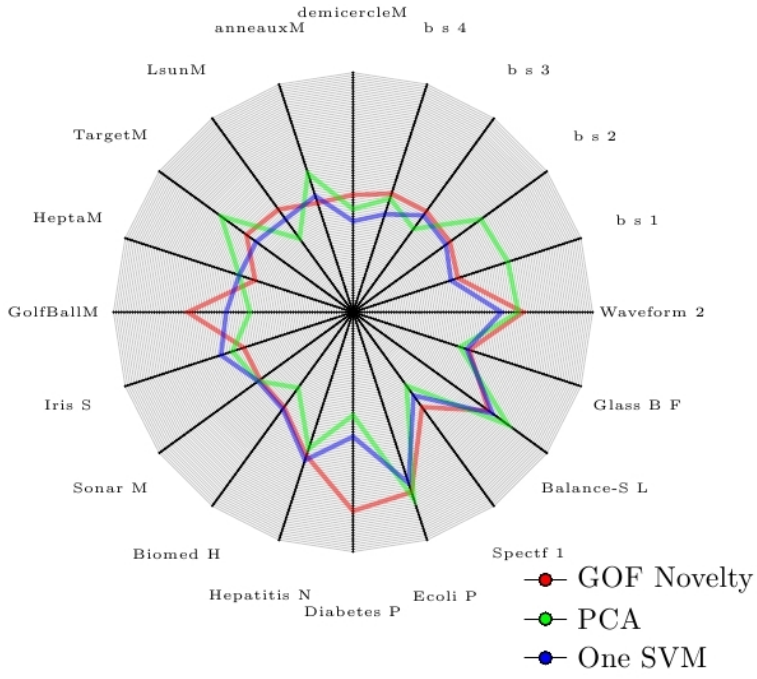
8: end if

9: end for

Threshold peut varier selon les bases (σ par défaut).



(a) Rappel



(b) AUC

Limites de GOF-Novelty

- Choix de la taille de la carte ;
- Choix du seuil de nouveautés.

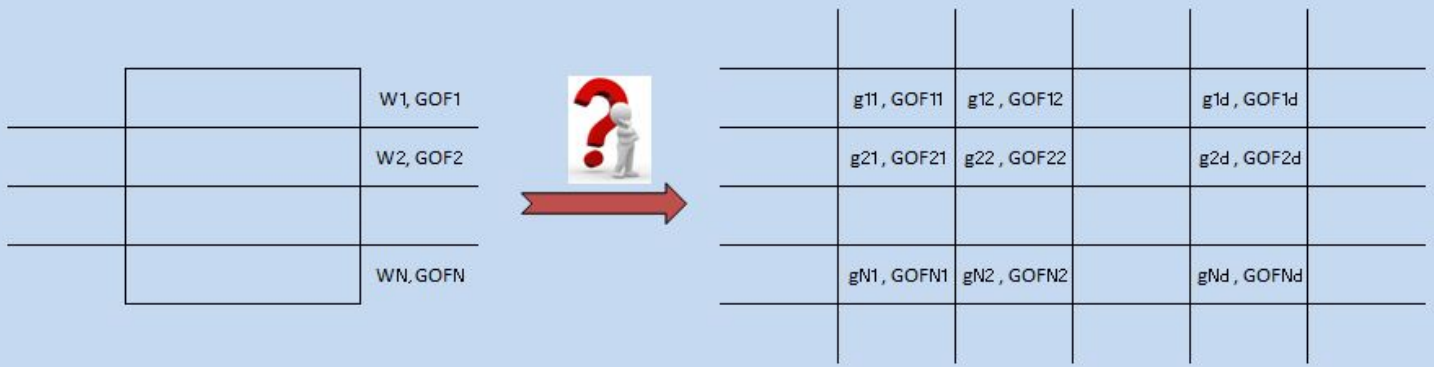
Bilan GOF

- Nouveau score pour la détection des "groupes-outliers" en utilisant les cartes topologiques ;
- Application à la détection de nouveautés.
- Application réelles :
  - Détection de fraudes ;
  - Surveillance des maladies ;
  - Astronomie.

Bilan GOF

- Nouveau score pour la détection des "groupes-outliers" en utilisant les cartes topologiques ;
- Application à la détection de nouveautés.
- Application réelles :
  - Détection de fraudes ;
  - Surveillance des maladies ;
  - Astronomie.

Détection des "blocs outliers"



Perspectives

1. **Flux de données massives :** estimation du seuil de passage d'un "groupe-outlier" à un "cluster-normal".
2. **Données binaires et mixtes :** GOF-SOM.
3. **Détection de nouveautés :** tester un autre seuil pour la détection de nouveautés ;